inovex

Big Data Projekte

Erfahrungsbericht aus der echten Welt

Nico Kreiling

Karlsruhe, 15.9.2016

Nico Kreiling

Big Data Scientist

- Wirtschaftsingenieurswesen am KIT, Abschluss März 2016
- Erfahrungen aus Web-Development und Datenbankadministration
- Bei der Inovex seit Juni 2016

Agenda

- Ein Big Data Projekt
- Technologiestack
- Erfahrungen aus der Praxis

Ausgangssituation

- Drei verschiedene Datenkategorien aus dem Webtracking: Clicks, Views und Events
- Tägliches Datenaufkommen > 1 Mrd
- Regelbewirtschaftung durch ETL-Strecken
- Schaffen einer Analyseplattform für Data Scientists

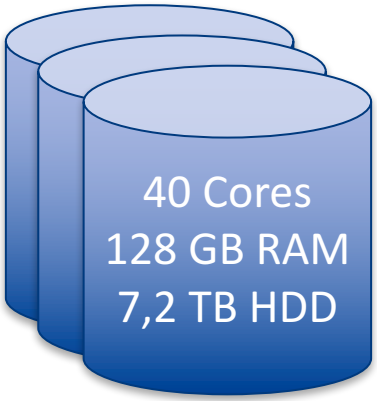
Big Data?

Was sind viele Daten Big-Data?

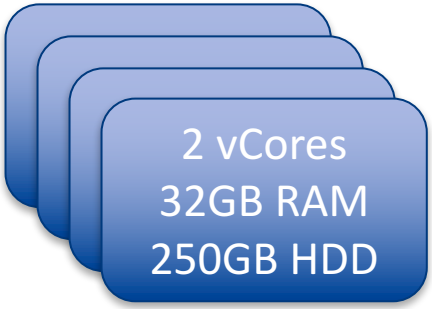
Volume	Velocity	Variety
• 6 TB	• ca. 1 Mrd/d • Batch-Verfahren (h)	• ca. 100 Attribute • Standardtypen

Big Data?

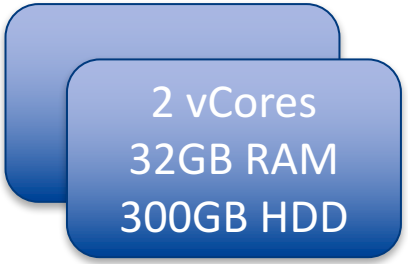
Mehr als auf einen Server passt



3 Server

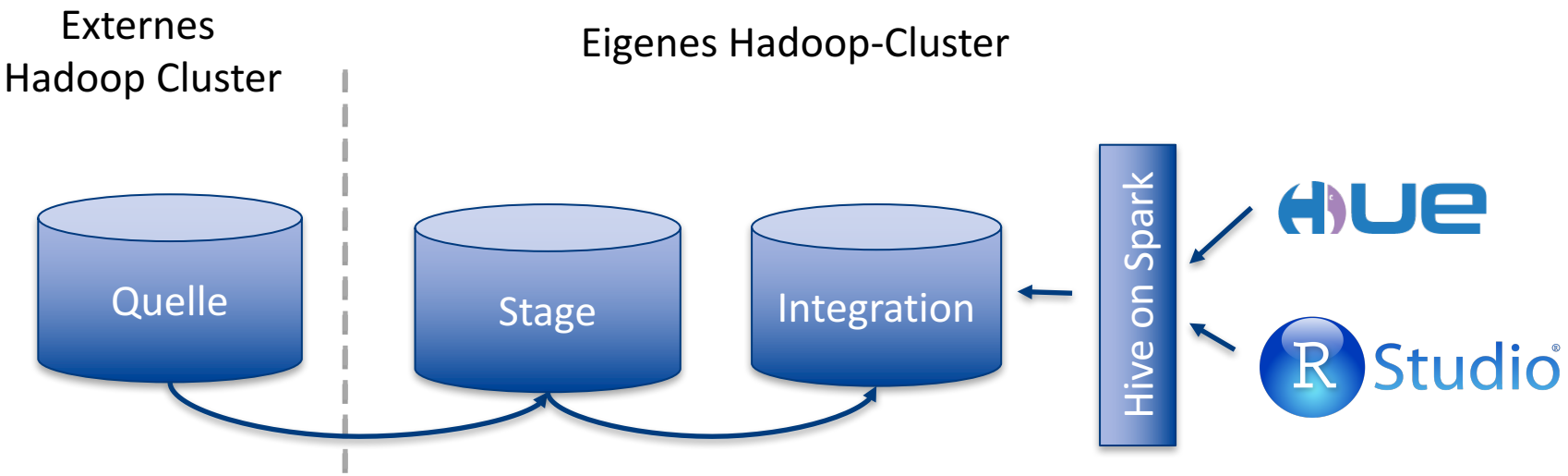


4 VMs Operativ



2 VMs Nutzer

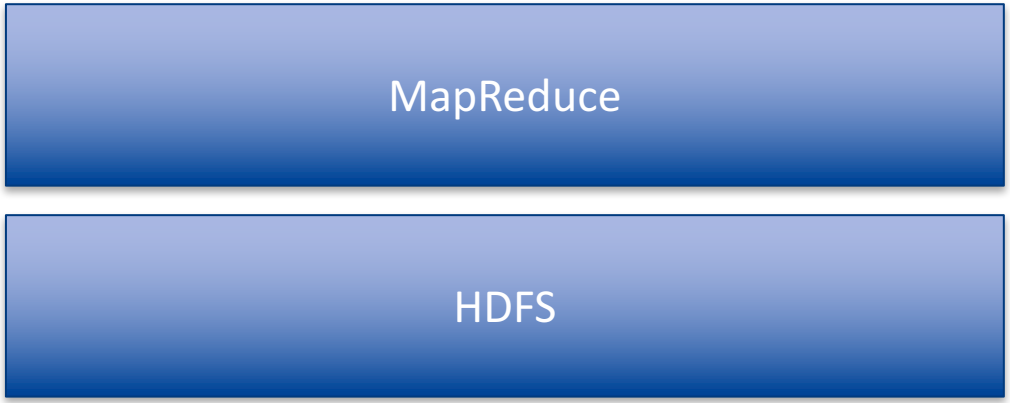
Beispiel einer ”Big Data”-Architektur



- Duplizieren
- Partitionieren

- Feature-Transformation
- Erweiterte Partitionierung
- Format-Mutation

Technologie-Stack: Das Fundament

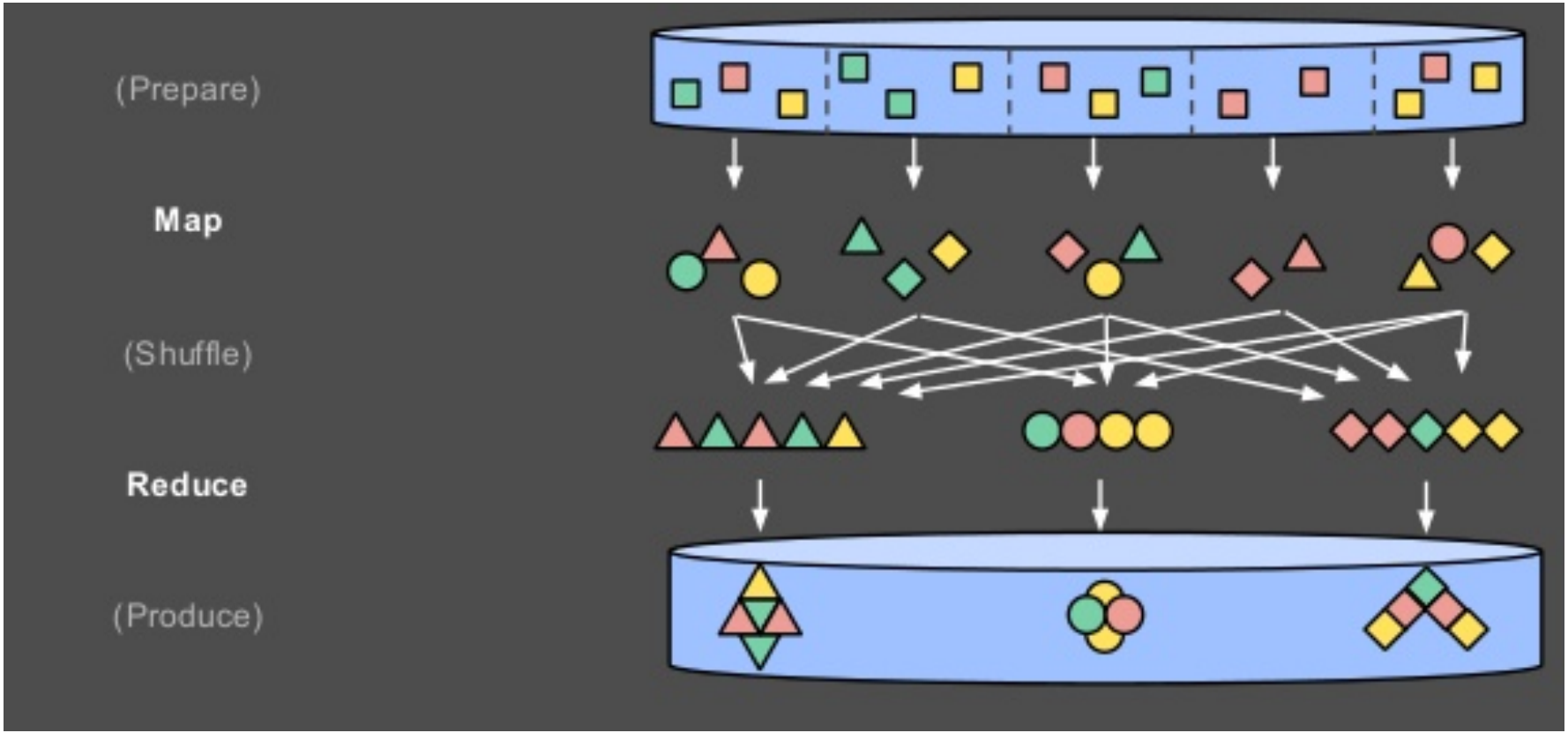


Hadoop Ökosystem

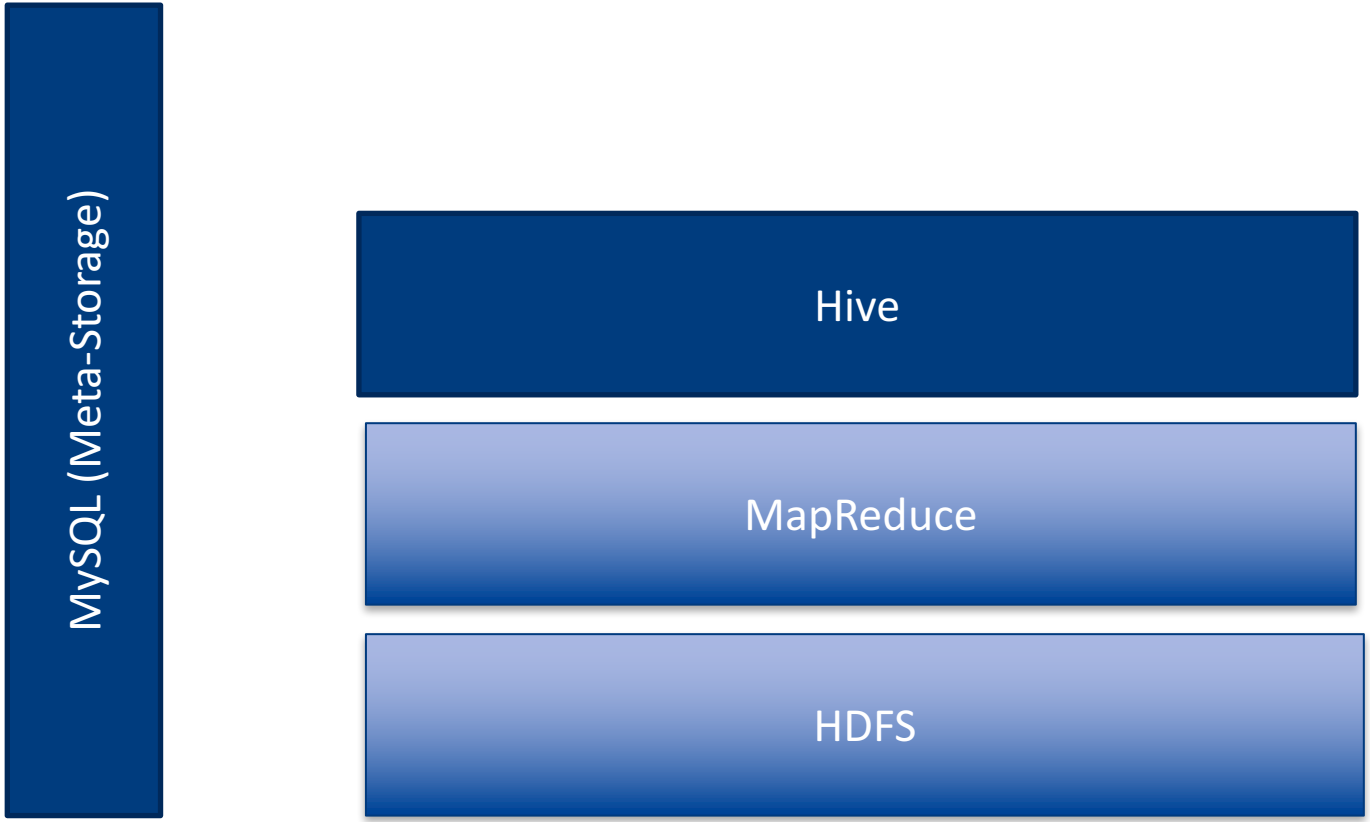
HDFS: Verteiltes Dateisystem, welches Dateiblöcke einer fixen Größe redundant speichert.

Map-Reduce: Parallelisierungsverfahren für verteilte Berechnungen

Map-Reduce Funktionsprinzip



Technologie-Stack: Datenaggregation



Hive

Hive: MapReduce-Generator basierend auf SQL Syntax

MySQL: Relationale Datenbank, wird hier als Metadaten-Speicher genutzt

Abfragensyntax: Hive vs MapReduce

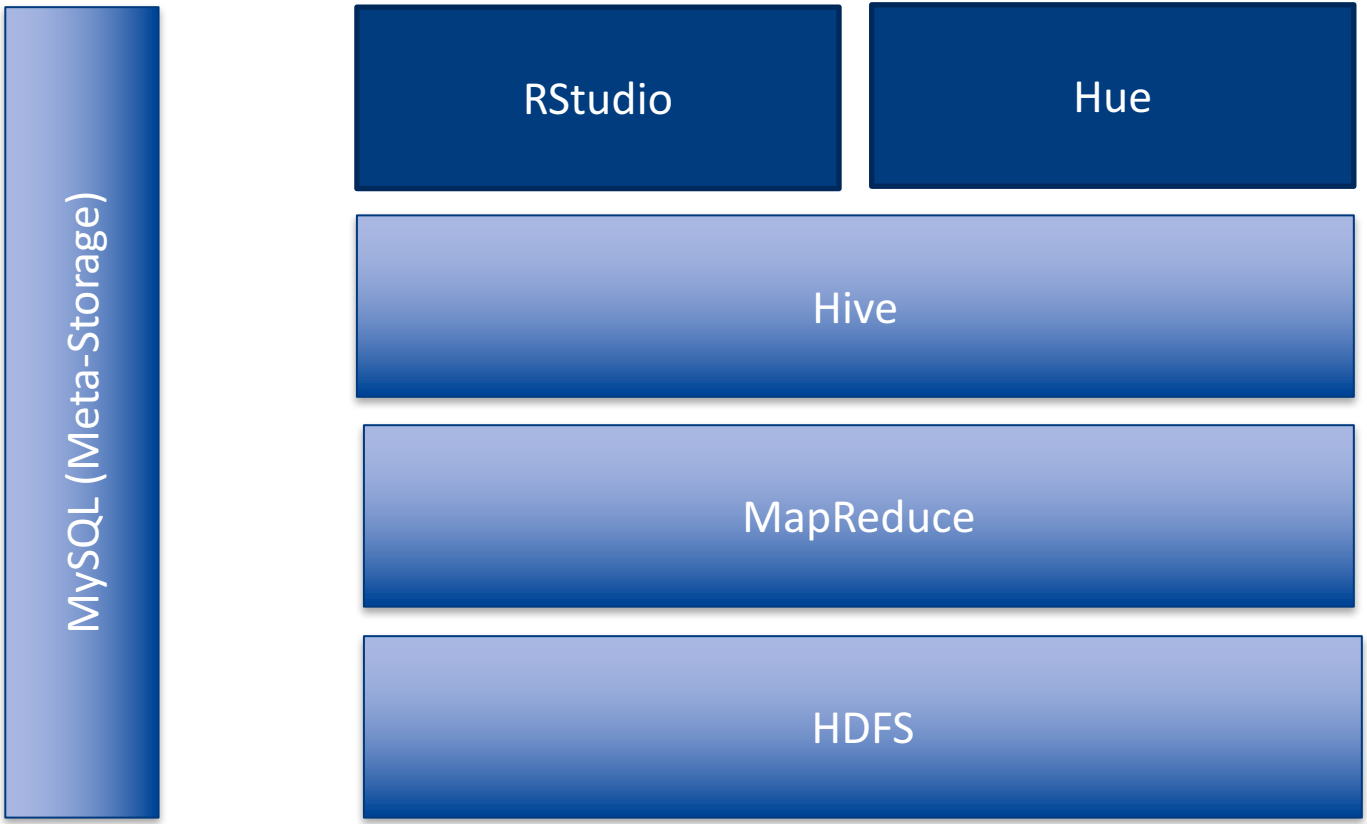
```java
package org.myorg;
import java.io.IOException;
import java.util.*;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.*;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

public class WordCount {
 public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(LongWritable key, Text value, Context context)
      throws IOException, InterruptedException {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
 }

 public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterable<IntWritable> values, Context context)
      throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        context.write(key, new IntWritable(sum));
    }
 }

 public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.waitForCompletion(true);
 }
}
```

```sql
CREATE TABLE docs (line STRING);

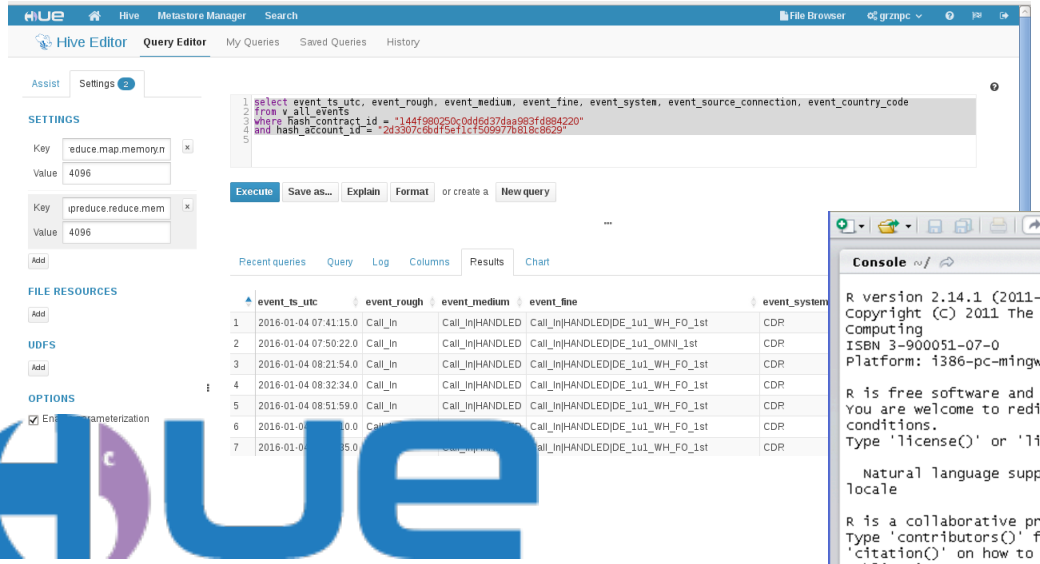
LOAD DATA INPATH 'docs' OVERWRITE INTO TABLE docs;

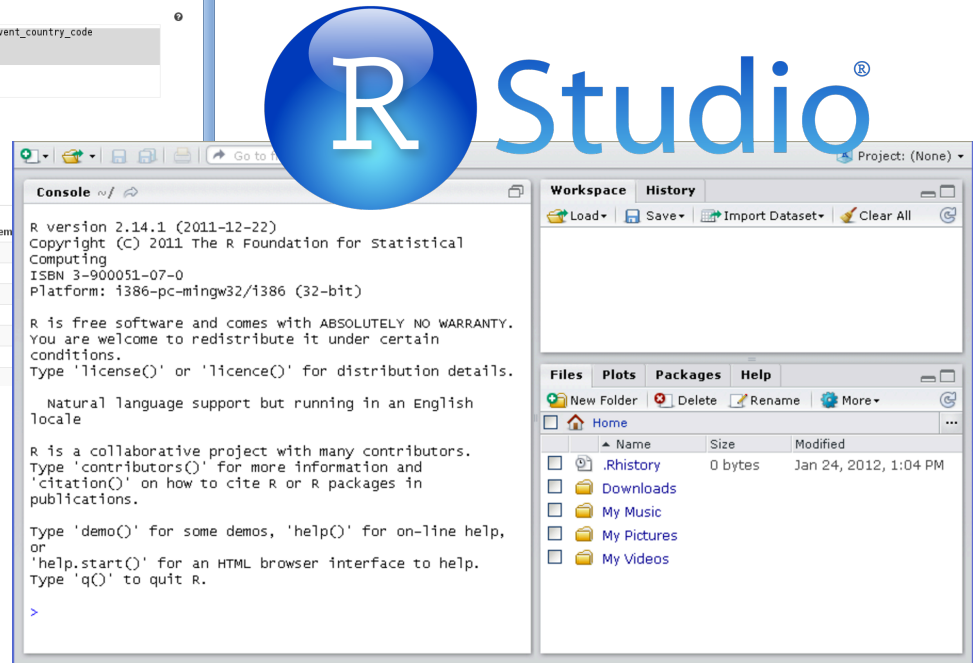
CREATE TABLE word_counts AS
  SELECT word, count(1) AS count
  FROM (
    SELECT explode(split(line, '\s')) AS word
    FROM docs) w
  GROUP BY word
  ORDER BY word;
```

Technologie-Stack: Auswertung

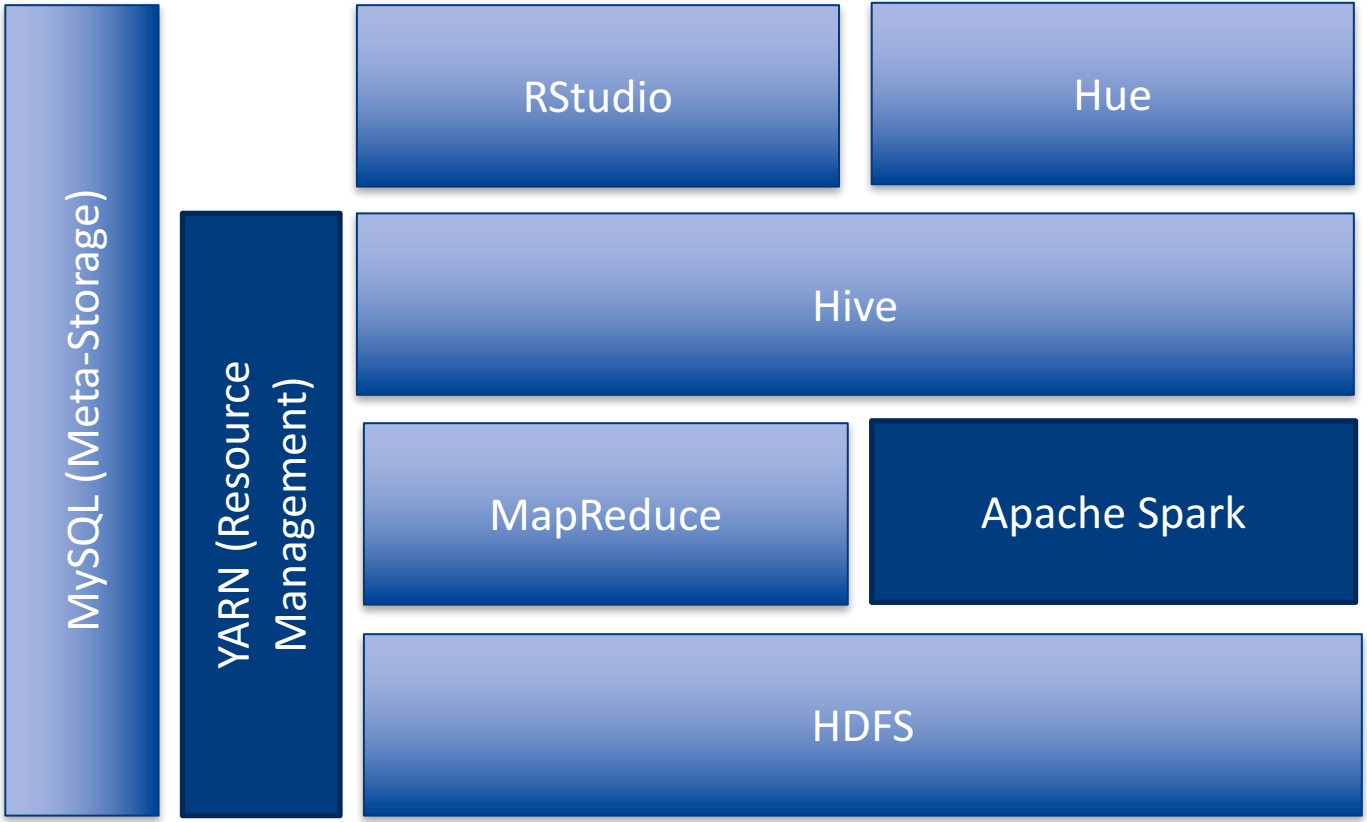


Auswertungstools: Hue & RStudio





Technologie-Stack

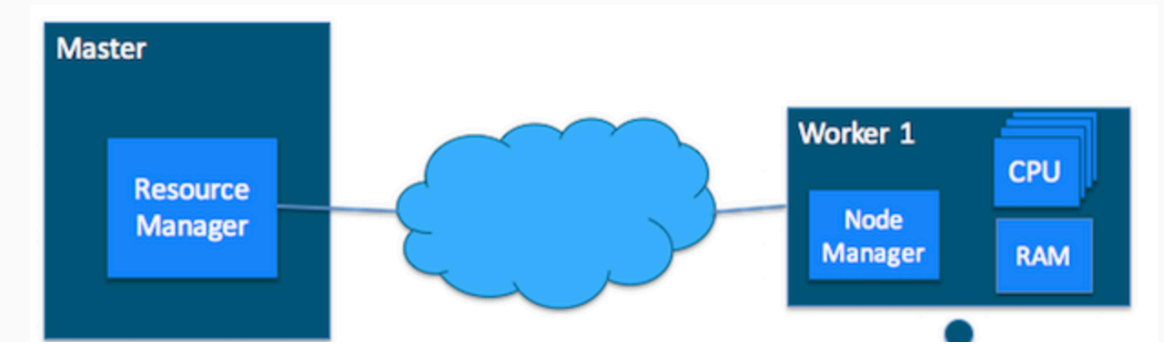


YARN

Yet Another Resource Negotiator

ResourceManager: Manages Ressources

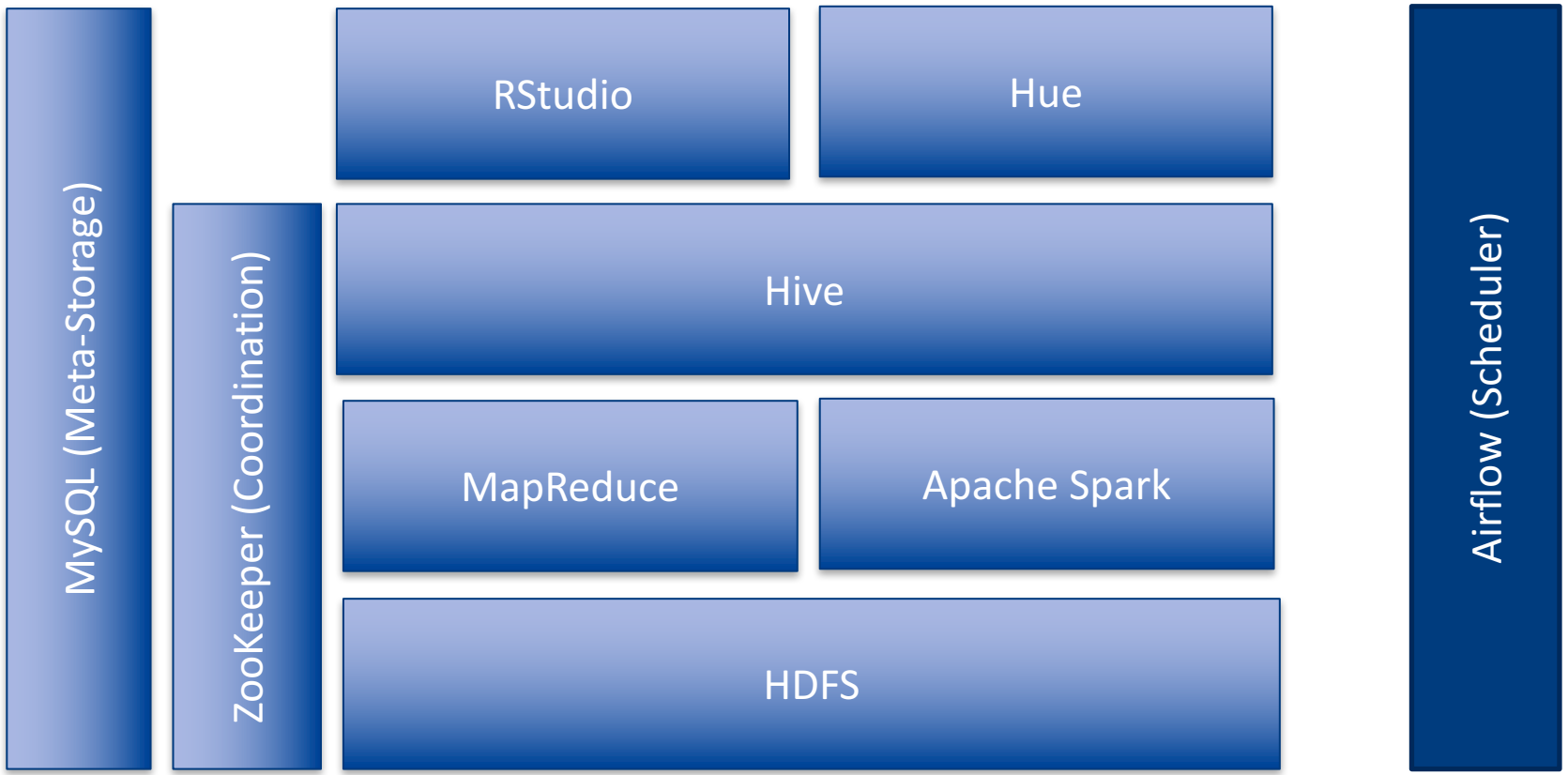
NodeManager: Handles the process execution



Spark

- In-Memory Berechnungen
- Graphstruktur (RDD)
- Lazy Evaluation

Technologie-Stack



Airflow

Workflow- und Job-Steuerung

Open Sourced von Airbnb

Praxiserfahrungen

- Initiales Setup ist ein wesentlicher Aufwand
- Low-Level IT-Wissen wichtig (shell, bash, OS-Befehle)

Echte Welt ≠ Hello World

Praxiserfahrungen – Lektion 1

- Mehr Daten > mehr Ressourcen >weniger Übersicht

- Struktur organisieren (Link-Page)
- TMUX

Diverses Technologiespektrum

Praxiserfahrungen – Lektion 2

Datenbanken:	Sprachen:	Hadoop:	Sonstiges
• MySQL • Sqlite • Oracle	• Python • Shell • SQL • YAML	• Map-Reduce • HDFS • Hive • Yarn	• Airflow • Ansible • Hue

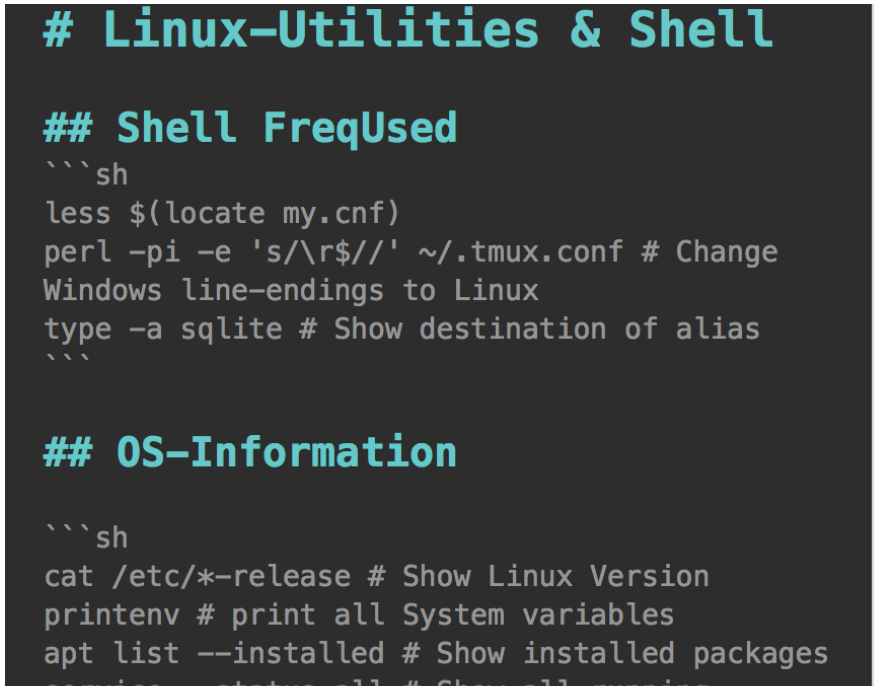
Noch diversere Auswahlmöglichkeiten

Praxiserfahrungen

MySQL PostgreSQL Oracle CouchDB MongoDB CouchDB Neo4j ElasticSearch Solr Redis Memcached Riak Dynamo Sqoop Accumulo Cassandra HBase Yarn Kubernetes Mesos Tensorflow TitanDB DockerSwarm Flink Storm Apache Streams Tez Spark MapReduce Impala Hive Pig Drill MicroStrategy Tableau Caravel Maven Pip Apt Airflow kaleidoscope Pandas Kerberos Kafka Flume Kibana NumPy Rstudio NiFi Solr Zeppelin Gearpump Greenplum

Lernen effizient gestalten

Praxiserfahrungen – Lektion 2



```
# Linux-Utilities & Shell

## Shell FreqUsed
```sh
less $(locate my.cnf)
perl -pi -e 's/\r$//' ~/.tmux.conf # Change
Windows line-endings to Linux
type -a sqlite # Show destination of alias
```

OS-Information

```sh
cat /etc/*-release # Show Linux Version
printenv # print all System variables
apt list --installed # Show installed packages
```

Linux-Utilities & Shell

Shell FreqUsed

```
less $(locate my.cnf)
perl -pi -e 's/\r$//' ~/.tmux.conf # Change
type -a sqlite # Show destination of alias
```

OS-Information

```
cat /etc/*-release # Show Linux Version
printenv # print all System variables
```
```

Testaufwand minimieren

Praxiserfahrungen – Lektion 3

- Wiederkehrende Kommandos in Skripte auslagern
- Kürzel in .bashrc definieren
- KeePass AutoType für Passwörter nutzen

Fazit

- Big-Data Projekte gibt es wirklich!
- Schnelle technische Weiterentwicklung und diverses Protfolio an Frameworks
- Erforderliches Skillset recht divers
- Viele Möglichkeiten sich zu verwirklichen

Vielen Dank

Nico Kreiling
Big Data Scientist

inovex GmbH
Ludwig-Erhard-Allee 6
76131 Karlsruhe

nico.kreiling@inovex.de
0173 3181 166